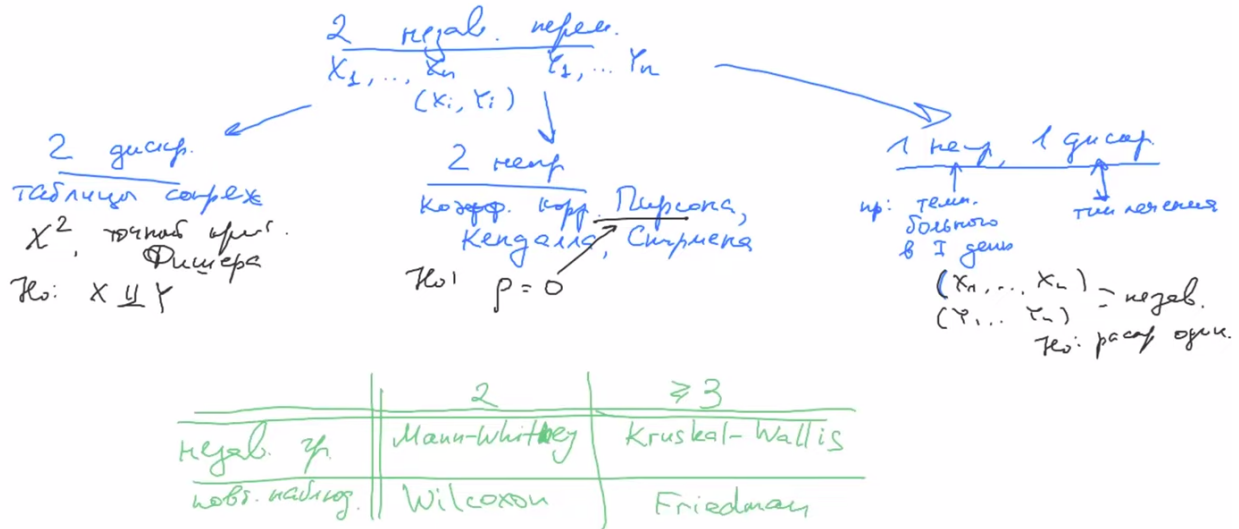


Занятие 9

Последние два занятия занимались вопросом того, как проверить, что две переменные независимы. Возможно три случая:



В третьем случае, когда одна переменная непрерывна, а другая дискретна, методы делятся на два типа: для независимых групп и для повторных наблюдений. Чтобы определить, с каким случаем мы имеем дело, надо задаться вопросом, можем ли мы посчитать попарные разности

Сегодня обсудим тесты Краскала-Уоллиса и Фридмана

Kruskal-Wallis

Пример

Many independent samples

Table 6.1 Half-Time of Mucociliary Clearance (h)

Normal subjects	Subjects with	
	Obstructive airways disease	Asbestosis
2.9 (8)	3.8 (13)	2.8 (7)
3.0 (9)	2.7 (6)	3.4 (11)
2.5 (4)	4.0 (14)	3.7 (12)
2.6 (5)	2.4 (3)	2.2 (2)
3.2 (10)		2.0 (1)
$R_1 = 36$	$R_2 = 36$	$R_3 = 33$

Source: M. L. Thomson and M. D. Short (1969).

Формализация

Несколько (≥ 3) независимых групп.

($k \geq 3$)
Несколько незав. групп

1	2	...	k
x_{11}	x_{12}		x_{1k}
	\vdots		
$x_{n_1 1}$	$x_{n_2 2}$		$x_{n_k k}$

Модель выглядит так:

$$x_{ij} = \Delta + \Delta_j + \varepsilon_{ij}$$

где

- Δ - общая медиана
- Δ_j - медиана по группе
- ε_{ij} - шум

Гипотеза:

$$\begin{aligned} \mathcal{H}_0 : & \Delta_1 = \dots = \Delta_k \\ \mathcal{H}_1 : & \text{else} \end{aligned}$$

В теории, не должно быть повторяющихся данных. Но в реальности часто такое случается. В этом случае берут средний ранг: если два элемента равны и имеют ранги 3 и 4, то обоим дают средний ранг - 3.5. Это чисто прикладной подход, не основанный на теории.

Пусть R_{ij} - ранг в общей выборке. Посчитаем средний ранг по группе (по столбцу):

$$R_j = \frac{1}{n_j} \sum_i R_{ij}$$

Теорема. Если выполнена \mathcal{H}_0 , то

$$\frac{12}{N(N+1)} \sum_{j=1}^k n_j \left(R_j - \frac{N+1}{2} \right)^2 \rightarrow \chi_{R-1}^2$$

По сути этот тест является обобщением теста Манни-Уитни

Связь с ANOVA

(aka однофакторный дисперсионный анализ)

Посчитаем *изменчивость по всей совокупности*:

$$\sum \sum (X_{ij} - X_{..})^2$$

где

- $X_{..}$ - среднее по всей выборке

Преобразуем:

$$\sum \sum (X_{ij} - X_{..})^2 = \underbrace{\sum_{j=1}^k \sum_{i=1}^{n_j} (X_{ij} - X_j)^2}_{\text{изменчивость внутри группы}} + \underbrace{\sum_{j=1}^k n_j (X_j - X_{..})^2}_{\text{изменчивость между группами}}$$

Это следует из теоремы Гюйгенса-Штейнера. Обозначим вещи выше:

$$V_{tot} = V_{int} + V_{out}$$

Теорема. Пусть $X_{ij} \sim N(\mu_j, \sigma^2)$. Если μ_1, \dots, μ_n , то $V_{int} \sim \chi_{N-k}^2$, $V_{out} \sim \chi_{k-1}^2$

Предположение о равенстве дисперсий довольно сильное, но есть тесты (критерий Баффлета), которые позволяют это проверить

Можно одновременно проверить оба условия, можно проверить следующее:

$$\frac{\frac{1}{k-1} V_{out}}{\frac{1}{N-k} V_{int}} \sim F_{k-1, N-k} - \text{распределение Фишера}$$

На основе этого работает метод ANOVA

Попробуем применить ANOVA для R_{ij} :

$$V_{tot} = \sum \sum (R_{ij} - R_{..})^2 = \sum_{i=1}^N \left(i - \frac{N+1}{2} \right)^2$$

Эта штука не зависит от выборки, а кроме того совпадает с выражением из асимптотической теоремы. Преобразуем:

$$\frac{\frac{1}{k-1} V_{out}}{\frac{1}{N-k} (V_{tot} - V_{out})} \sim F_{k-1, N-k}$$

Получаем, что метод Краскала-Уоллиса и ANOVA базируются на одном и том же утверждении. Поэтому критерий КУ называют критерием непараметрического дисперсионного анализа.

Критерий Фридмана

Пример

Пример из бейсбола. Надо пробежать от home plate до second base мимо first base. Есть три стратегии, надо понять какая лучше.

Rounding the first base in baseball

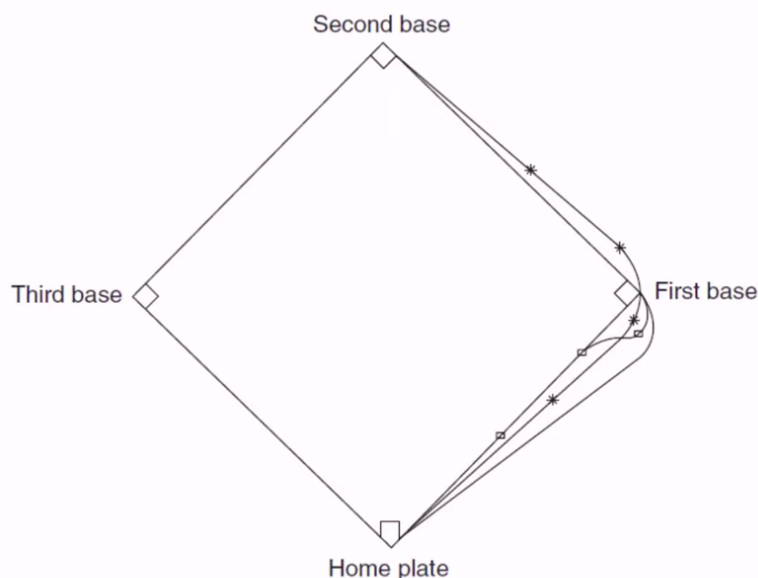


Figure 7.1 Three methods of rounding first base: ◇ path = round out method, * path = narrow angle method, solid path = wide angle method.

Результаты:

Rounding times

Table 7.1 Rounding-First-Base Times

Players	Methods		
	Round out	Narrow angle	Wide Angle
1	5.40 (1)	5.50 (2)	5.55 (3)
2	5.85 (3)	5.70 (1)	5.75 (2)
3	5.20 (1)	5.60 (3)	5.50 (2)
4	5.55 (3)	5.50 (2)	5.40 (1)
5	5.90 (3)	5.85 (2)	5.70 (1)
6	5.45 (1)	5.55 (2)	5.60 (3)
7	5.40 (2.5)	5.40 (2.5)	5.35 (1)
8	5.45 (2)	5.50 (3)	5.35 (1)
9	5.25 (3)	5.15 (2)	5.00 (1)
10	5.85 (3)	5.80 (2)	5.70 (1)
11	5.25 (3)	5.20 (2)	5.10 (1)
12	5.65 (3)	5.55 (2)	5.45 (1)
13	5.60 (3)	5.35 (1)	5.45 (2)
14	5.05 (3)	5.00 (2)	4.95 (1)
15	5.50 (2.5)	5.50 (2.5)	5.40 (1)
16	5.45 (1)	5.55 (3)	5.50 (2)
17	5.55 (2.5)	5.55 (2.5)	5.35 (1)
18	5.45 (1)	5.50 (2)	5.55 (3)
19	5.50 (3)	5.45 (2)	5.25 (1)
20	5.65 (3)	5.60 (2)	5.40 (1)
21	5.70 (3)	5.65 (2)	5.55 (1)
22	6.30 (2.5)	6.30 (2.5)	6.25 (1)
	$R_1 = 53$	$R_2 = 47$	$R_3 = 32$

Source: W. F. Woodward (1970).

Наблюдения парные (одни и те же игроки)

Формализация

Friedman's test

treatment block	1	2	...	k
1	$X_{11}^{(1)}$	$X_{12}^{(1)}$		$X_{1k}^{(1)}$
	$X_{11}^{(C_{11})}$	$X_{12}^{(C_{12})}$		$X_{1k}^{(C_{1k})}$
2	$X_{21}^{(1)}$	$X_{22}^{(1)}$		$X_{2k}^{(1)}$
	$X_{21}^{(C_{11})}$	$X_{22}^{(C_{12})}$		$X_{2k}^{(C_{1k})}$
...				
n	$X_{n1}^{(1)}$	$X_{n2}^{(1)}$		$X_{nk}^{(1)}$
	$X_{n1}^{(C_{11})}$	$X_{n2}^{(C_{12})}$		$X_{nk}^{(C_{1k})}$

Модель:

$$x_{ij}^{(k)} = \underset{\text{медиана}}{\theta} + \underset{\substack{\text{эффект блока} \\ \text{(способности игрока)}}}{\alpha_i} + \underset{\substack{\text{эффект обработки} \\ \text{(траектория)}}}{\beta_i} + \varepsilon_{ij}^{(k)}, \quad k = \overline{1, C_{ij}}$$

Гипотеза:

$$\mathcal{H}_0: \beta_1 = \dots = \beta_k$$

Обозначения:

$$r_{ij}^{(k)} = \text{rank}(X_{ij}^{(k)}) - \text{внутри } i\text{-го блока}$$

$$R_{.j} = \frac{1}{C_{11} + \dots + C_{n1}} \sum_{i=1}^n \sum_{j=1}^{C_{ij}} r_{ij}^{(k)}$$

Асимптотическая теорема

Если \mathcal{H}_0 верна, то

$$\frac{12n}{k(k+1)} \sum_{j=1}^k \left(R_{.j} - \frac{k+1}{2} \right)^2 \rightarrow \chi_{k-1}^2$$

Регрессия

Предыстория

В 1885 Гальтон измерил рост родителей X и рост детей Y . Он обнаружил, что $Y - \bar{Y} = \frac{2}{3}(X - \bar{X})$

Вообще термин регрессия немного депрессивный и намекает на уменьшение/затухание. В советской литературе этот термин был заменен словом прогрессия

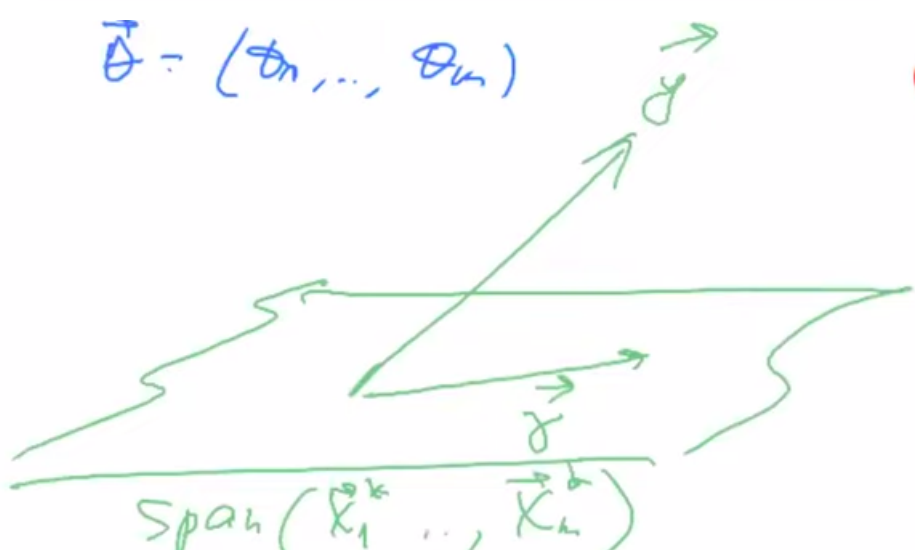
Линейная регрессия

Модель:

$$(\vec{X}_i, Y_i), i = \overline{1, n}, \quad \dim X_i = m$$

$$Y_i = (\vec{X}_i, \vec{\theta}) + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2)$$

Записываем в виде матрицы:

$$\begin{aligned} Y_i &= (\vec{X}_i, \vec{\theta}) + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2) \\ \underbrace{\begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}}_{\vec{y}} &= \underbrace{\begin{pmatrix} X_{11} & \dots & X_{1m} \\ \vdots & & \vdots \\ X_{n1} & \dots & X_{nm} \end{pmatrix}}_{\text{design matrix}} \underbrace{\begin{pmatrix} \theta_1 \\ \vdots \\ \theta_m \end{pmatrix}}_{\vec{\theta}} + \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix} \\ &\rightarrow \theta_1 \cdot \underbrace{\begin{pmatrix} X_{11} \\ \vdots \\ X_{n1} \end{pmatrix}}_{\vec{X}_1^*} + \dots + \theta_m \cdot \underbrace{\begin{pmatrix} X_{1m} \\ \vdots \\ X_{nm} \end{pmatrix}}_{\vec{X}_m^*} \\ &\quad \parallel \vec{\theta} \\ \vec{\theta} &= (\theta_1, \dots, \theta_m) \end{aligned}$$


Как решать? МНК

$$\begin{aligned}
 \text{МНК: } \sum_{i=1}^n (y_i - \sum_{j=1}^k x_{ij} \theta_j)^2 &= \| \vec{y} - X \vec{\theta} \|^2 \xrightarrow{\theta} \min \\
 \arg \min_{\vec{\theta} \in L_0} \| \vec{y} - \vec{\theta} \|^2, \quad \vec{y} - X \vec{\theta} \perp L_0 &\Leftrightarrow (\vec{y} - X \vec{\theta}, \vec{x}_i^*) = 0 \quad \forall i \\
 X^T (\vec{y} - X \vec{\theta}) &= 0 \\
 X^T \vec{y} - X^T X \vec{\theta} &= 0 \\
 \boxed{\vec{\theta} = (X^T X)^{-1} X^T \vec{y}}
 \end{aligned}$$

$$\begin{aligned}
 y_i &= (\vec{x}_i, \vec{\theta}) + \varepsilon_i, \\
 \vec{\theta} &= (X^T X)^{-1} X^T \vec{y}
 \end{aligned}$$

По сути наши предсказания являются л.к. у-ков:

$$\begin{pmatrix} \hat{y}_1 \\ \vdots \\ \hat{y}_n \end{pmatrix} = H \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}; \quad H = X^T (X^T X)^{-1} X^T$$

↑
linear smoothers

Пример

А где тут матстатистика? На семинарах будем работать с датасетом mtcars. Будем решать следующие проблемы:

- Какие признаки лучше брать для предсказания?
- Почему построенная модель лучше тривиальной (прогнозирующей константу)?

Проблема уменьшения размерности



Пусть $L_1 \subset L_0$

$$\vec{Y} = \vec{\gamma} + \vec{\varepsilon}, \quad \vec{\gamma} \in L_0, \quad \varepsilon_i \in N(0, \sigma^2)$$

Теорема. Пусть

$$\mathcal{H}_0: \vec{\gamma} \in L_1 \subset L_0$$

Если \mathcal{H}_0 верна, то

$$\frac{\frac{1}{n_0 - n_1} \|\text{proj}_{L_0} \vec{y} - \text{proj}_{L_1} \vec{y}\|^2}{\frac{1}{n - n_0} \|\vec{y} - \text{proj}_{L_0} \vec{y}\|^2} \sim F_{n_0 - n_1, n - n_0}$$

Связь с линейной корреляцией

Вспомним коэффициент R^2 :

$$R^2 = \frac{\|\text{proj}_{L_0} \vec{y} - \text{proj}_{L_1} \vec{y}\|^2}{\|\vec{y} - \text{proj}_{L_0} \vec{y}\|^2} = \left\{ \text{pifagorean theorem} \right\} = \frac{\|\text{proj}_{L_0} \vec{y} - \text{proj}_{L_1} \vec{y}\|^2}{\|\text{proj}_{L_0} \vec{y} - \text{proj}_{L_1} \vec{y}\|^2 + \|\vec{y} - \text{proj}_{L_0} \vec{y}\|^2}$$

Тогда можно переписать отверждение теоремы:

$$\frac{R^2}{1 - R^2} \cdot \frac{n - n_0}{n_0 - n_1} \sim F_{n_0 - n_1, n - n_0}$$

Это похоже на теорему для коэффициента корреляции Пирсона. Возьмем в качестве L_0 пространство, натянутое на два вектора:

$$\begin{aligned} L_0 &= \text{span}(x_1^*, x_2^*), & y &= a + bx + \varepsilon \\ L_1 &= \text{span}(x_1^*), & y &= a + \varepsilon \end{aligned}$$

Тогда получаем

$$\frac{R^2}{1 - R^2} \cdot \frac{n - 2}{1} \sim F_{1, n-2} - \text{распределение Фишера}$$

$$F_{k,n} = \frac{\frac{1}{k}(\xi_1^2 + \dots + \xi_k^2)}{\frac{1}{n}(\xi_1^2 + \dots + \xi_n^2)}$$

$$\frac{R}{\sqrt{1 - R^2}} \sqrt{n - 2} \sim t_{n-2} - \text{распределение Стьюдента}$$

Почему совпадает???

Связь коэффициента корреляции Пирсона и R^2

Выборка:

$$(x_1, y_1), \dots, (x_n, y_n)$$

Коэффициент корреляции:

$$\text{emp.cor}(x, y) = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2} \sqrt{\sum (y_i - \bar{y})^2}}$$

Регрессионная зависимость:

$$y_i = a + bx_i + \varepsilon_i$$

С помощью МНК получаем $\hat{a}, \hat{b} \Rightarrow \hat{y}_i = \hat{a} + \hat{b}x_i$

Утверждение 1.

$$|\text{emp.cor.}(a + bx, y)| = |\text{emp.cor.}(x, y)|$$

Доказательство

$$\text{л.ч.} = \frac{\sum (\cancel{a} + bx_i - (\cancel{a} + b\bar{x})) (y_i - \bar{y})}{\sqrt{\sum (\cancel{a} + bx_i - \cancel{a} - b\bar{x})^2} \cdot \sqrt{\sum (y_i - \bar{y})^2}} = \frac{b}{|b|} \text{emp.cor.}(x, y)$$

"если 1, если -1" \square

Утверждение 2.

$$(\text{emp.cor.}(\hat{y}, y))^2 = R^2$$

Доказательство

Запишем левую и правую часть. Явно посчитаем проекции:

усл. 2
for

$$\left(\text{emp.cor.}(\hat{y}, y) \right)^2 = R^2$$

$$\text{л.ч.} = \frac{\left(\sum (y_i - \text{mean}(y)) (\hat{y}_i - \text{mean}(\hat{y})) \right)^2}{\sum (y_i - \text{mean}(y))^2 \cdot \sum (\hat{y}_i - \text{mean}(\hat{y}))^2}$$

$$\text{пр.ч.} = \frac{\|P_L \hat{y} - P_L \bar{y}\|^2}{\|y - P_L \bar{y}\|^2}$$

$P_L \bar{y} = \arg \min_c \sum (y_i - c)^2 = \text{mean}(y)$
 $P_L \hat{y} = \arg \min_{\alpha, \beta} \sum (y_i - \alpha - \beta x_i)^2 = (\hat{\alpha}, \hat{\beta})$

Значит, правая часть равна:

$$\text{пр.ч.} = \frac{\sum (\hat{y}_i - \text{mean}(\hat{y}))^2}{\sum (y_i - \text{mean}(y))^2}$$

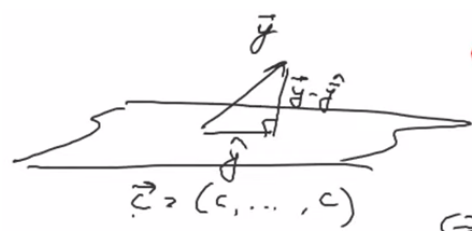
Заметим следующее:

Если $\text{mean}(y) = \text{mean}(\hat{y}) = c$, то л.ч. = пр.ч.

$$\frac{(\sum (y_i - c) (\hat{y}_i - c))^2}{\sum (y_i - c)^2 \sum (\hat{y}_i - c)^2} = \frac{\sum (\hat{y}_i - c)^2}{\sum (y_i - c)^2}$$

$$(\sum (y_i - c) (\hat{y}_i - c))^2 = (\sum (\hat{y}_i - c))^2$$

Последнее тождество верно, что можно увидеть из картинки:



$$\vec{c} = (c, \dots, c)$$

$$(\vec{y} - \vec{c}, \vec{y} - \vec{c}) = \|\vec{y} - \vec{c}\|^2$$

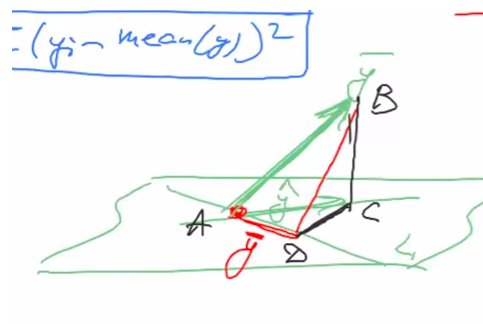
$$(\vec{y} - \vec{y} - \vec{c}, \vec{y} - \vec{y} - \vec{c}) = \|\vec{y} - \vec{y} - \vec{c}\|^2 = \|\vec{y} - \vec{c}\|^2$$

$$\Rightarrow ((\vec{y} - \vec{y} + \vec{y} - \vec{c}, \vec{y} - \vec{c})) = \|\vec{y} - \vec{c}\|^2$$

$$= ((\vec{y} - \vec{y}, \vec{y} - \vec{c}) + \|\vec{y} - \vec{c}\|^2) = \|\vec{y} - \vec{c}\|^2$$

тогда $a = \hat{a}$, $b = \hat{b}$

Осталось доказать, что $\text{mean}(y) = \text{mean}(\hat{y})$. Вспоминаем школьную теорему о трех перпендикулярах:



$$= (y_i - \text{mean}(y))^2$$

$$AD \perp BC \text{ (т.к. } BC \perp L_0) \left\{ \begin{array}{l} AD \perp BA \text{ (т.к. } BA \perp L_1) \end{array} \right\} \Rightarrow$$

$$\Rightarrow AD \perp CA \Rightarrow$$

$$\Rightarrow \hat{y} = \bar{y}$$

$$\text{mean}(\hat{y}) = \text{mean}(y). \quad \square$$

Возьмем $a = \hat{a}$, $b = \hat{b}$. Тогда

$$\text{emp.cor.}(\hat{y}, y) = |\text{emp.cor.}(x, y)|$$

Значит,

$$R^2 = (\text{emp.cor.}(x, y))^2$$

Если утверждение 2 верно, то мы можем использовать теорему, выписанную ранее:

$$\mathcal{H}_0 : \gamma \in L_1 \Rightarrow \frac{R}{\sqrt{1 - R^2}} \sqrt{n - 2} \sim t_{n-2}$$

Условие $\gamma \in L_1$ эквивалентно равенству нулю линейной корреляции:

$$\rho = 0 \Leftrightarrow \mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y] \Leftrightarrow \{Y = \alpha + \beta X + \varepsilon\} \Leftrightarrow \alpha \mathbb{E}[X] + \beta \mathbb{E}[X^2] = \mathbb{E}[X](\alpha + \beta \mathbb{E}[X]) \Leftrightarrow$$

$$\Leftrightarrow \beta \text{Var}[X] = 0 \Leftrightarrow \beta = 0 \Leftrightarrow \gamma \in L_1$$

Поэтому анализ коэффициента корреляции Пирсона полностью эквивалентен анализу коэффициента R^2 в соответствующей модели.

Обобщенные линейные модели

Пусть есть экспоненциальное распределение

$$p(x, v) = g(x) \exp\{xv - d(v)\}$$

и $Y \sim p(x, v)$. Пусть имеется выборка $(\bar{X}_0, Y_1), \dots, (\bar{X}_n, Y_n)$. Будем предполагать, что

$$Y_i \sim p(x, (\bar{x}_i, \beta))$$

Пример: логистическая регрессия

$$Y \sim \begin{cases} 1, & \theta \\ 0, & 1 - \theta \end{cases}$$

Плотность:

$$p(x, \theta) = \theta^x (1 - \theta)^{1-x} = \exp\{x \log \theta + (1-x) \log(1 - \theta)\}$$

Таким образом,

$$Y_i \sim \exp\{x \langle y_i, \beta \rangle - d(\langle y_i, \beta \rangle)\}, \quad d(v) = \log(1 + e^v)$$

Чтобы найти β используем метод максимального правдоподобия:

$$\prod_{i=1}^n p(x_i, (y_i, \beta)) = \prod_{i=1}^n \exp\{x_i \langle y_i, \beta \rangle - d(\langle y_i, \beta \rangle)\} \rightarrow \max_{\beta}$$

Анализ качества логистической регрессии

Кривая ROC-AUC

Бинаризуем по порогу:

логист. регр:

$$Y_i \sim p(x, (\bar{x}_i, \beta)) \rightarrow \hat{\beta}$$

$$p(x, \theta) = e^{x\theta - d(\theta)}$$

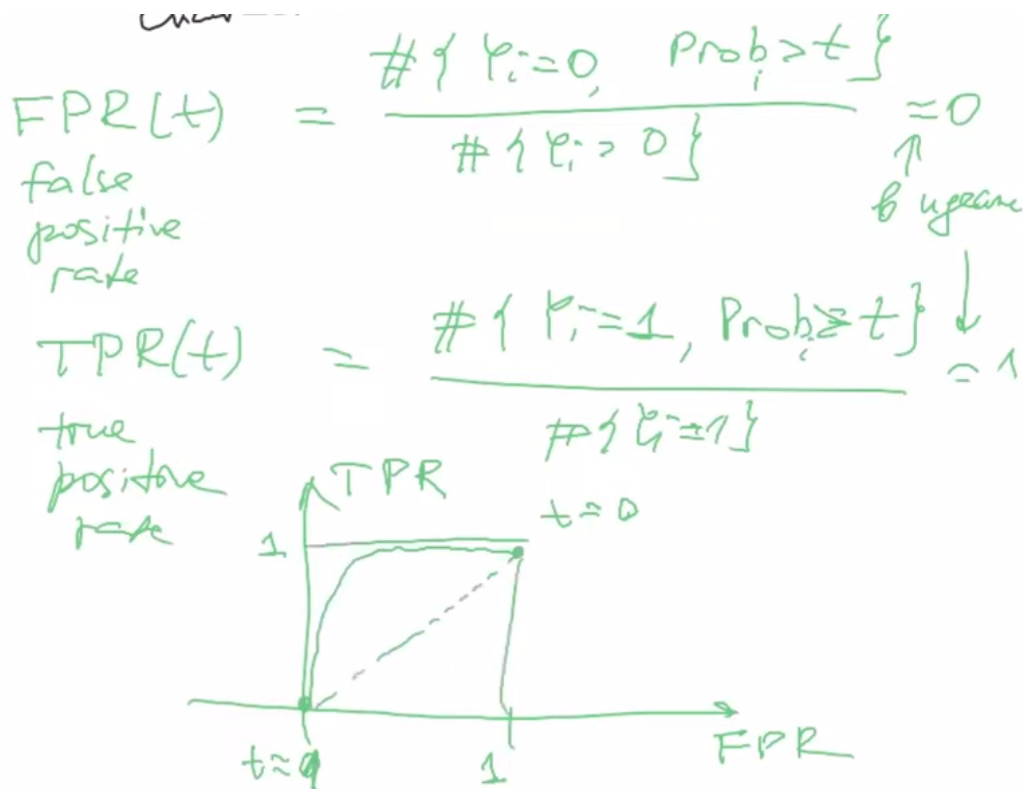
$$\theta = \frac{e^v}{1+e^v} \Rightarrow \hat{\theta}_i = \frac{e^{\langle \bar{x}_i, \hat{\beta} \rangle}}{1 + e^{\langle \bar{x}_i, \hat{\beta} \rangle}}$$

Y	Prob
0	0.67
1	0.82
0	0.17
0	0.87

$\text{Prob} > t \Rightarrow Y = 1$

Подбор порога, вообще говоря, довольно важен.

TPR и FPR:



Теорема Уилкса

- Null deviance:

$$\text{ND} = 2 \cdot \left(\text{LogLikelihood}(\text{saturated model}) - \text{LogLikelihood}(\text{null model}) \right) \rightarrow \chi^2_{n-1}$$

$Y_i \sim p(x, v_i)$
 $Y_i \sim p(x, v)$

- Residual deviance:

$$\text{RD} = 2 \cdot \left(\text{LogLikelihood}(\text{saturated model}) - \text{LogLikelihood}(\text{proposed model}) \right) \rightarrow \chi^2_{n-(m+1)}$$

$Y_i \sim p(x, v_i)$
 $Y_i \sim p(x, (\theta, x_i))$

Когда мы оценивали линейную модель, то по сути сравнивали proposed model и saturated model.

Чтобы оценить нашу модель просто посчитаем разность и применим статистический тест:

$$\text{ND} - \text{RD} = 2 \cdot \left(\text{LogLikelihood}(\text{proposed model}) - \text{LogLikelihood}(\text{null model}) \right) \rightarrow \chi^2_m$$

Weights of evidence

Пример: событие это уход клиента из банка.

Стоит ли использовать переменную "возраст клиента" для предсказания оттока?

Разобьем на бины:

bins	non-event	event
1	20	11
2	30	7
3	40	68
⋮		
K	18	61
	Σ_{ne}	Σ_e

Статистика:

$$WoE_i = \ln \left(\frac{ne_i / \Sigma_{ne}}{e_i / \Sigma_e} \right)$$

где i - номер бина. Information value:

$$IV = \sum_{i=1}^k \left(\underbrace{\frac{ne_i}{\Sigma_{ne}}}_{g_i} - \underbrace{\frac{e_i}{\Sigma_e}}_{h_i} \right) \cdot WoE_i = \sum (g_i - h_i) \ln(g_i/h_i) = \sum g_i \ln(g_i/h_i) + \sum h_i \ln(h_i/g_i) \geq 0$$

Причем равенство достигается, когда $h = g$. Соответственно, чем меньше IV, тем хуже переменная.

Таблица для определения качества:

IV < 0.02 - useless for prediction
0.02-0.1 - weak prediction
0.1-0.3 - medium
0.3-0.5 - strong
> 0.5 - too good to be true

Непараметрическая регрессия

Kernel regression

$$Y_i = r(X_i) + \varepsilon_i$$

$$\mathbb{E}[Y_i | X_i = x] = \mathbb{E}[r(X_i) | X_i = x] + \underbrace{\mathbb{E}[\varepsilon | X_i = x]}_0$$

$$\mathbb{E}[Y | X = x] = \int y p_{y|x}(y, x) dy = \int y \cdot \frac{p_{x,y}(x, y)}{p_x(x)} dy$$

Мы не знаем обе плотности в последние интеграле. Будем использовать ядерные оценки:

$$\hat{p}_x(x) = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{x_i - x}{h}\right)$$

$$\hat{p}_{(x,y)}(x, y) = \frac{1}{nh^2} \sum_{i=1}^n K\left(\frac{x_i - x}{h}\right) K\left(\frac{y_i - y}{h}\right)$$

Подставляем:

$$\hat{r}(x) = \frac{\int y \hat{p}_{(x,y)}(x,y) dy}{\hat{p}_x(x)} = \frac{\frac{1}{nh^2} \sum_{i=1}^n K\left(\frac{x_i-x}{h}\right) K\left(\frac{y_i-y}{h}\right)}{\frac{1}{nh} \sum_{i=1}^n K\left(\frac{x_i-x}{h}\right)} = \frac{\sum_{i=1}^n y_i \cdot K\left(\frac{x_i-x}{h}\right)}{\sum_{i=1}^n K\left(\frac{x_i-x}{h}\right)}$$

$$\int y K\left(\frac{x-y}{h}\right) dy = \left[\frac{y_i-y}{h} = u \right] = \int (y_i - uh) K(u) \cdot h du = y_i \cdot h \int K(u) du - h^2 \int u K(u) du = y_i \cdot h - 0$$

Итоговая оценка:

$$\hat{r}(x) = \frac{\sum_{i=1}^n y_i K\left(\frac{x_i-x}{h}\right)}{\sum_{i=1}^n K\left(\frac{x_i-x}{h}\right)} - \text{Nadaraya-Watson}$$

Утверждение. Оценка Надарая-Ватсона является решение оптимизационной задачи

$$\operatorname{argmin}_{\beta} \sum_{i=1}^n (y_i - \beta)^2 K\left(\frac{x - x_i}{h}\right)$$

Обобщение:

$$\operatorname{argmin}_{\beta_0, \beta_1} \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i) K\left(\frac{x - x_i}{h}\right)$$

Резюме

Регрессионные модели бывают двух типов - параметрические и непараметрические. Среди параметрических самая мощная - линейная регрессия, т.к. для нее существует много статистических тестов.

Линейная регрессия плотно связана с корреляцией Пирсона: они обе используют одну и ту же теорему.

Существуют обобщения линейных моделей. Логистическая регрессия одна из них.